OXFORD
UNIVERSITY PRESS

Archives
of
CLINICAL
NEUROPSYCHOLOGY

A Test Battery for Inner Speech Functions

Sharon Geva[1,2], Elizabeth A. Warburton[1,*]

[1]*Department of Clinical Neurosciences, University of Cambridge, R3 Neurosciences – Box 83, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK*
[2]*Current address: Cognitive Neuroscience and Neuropsychiatry Section, UCL Institute of Child Health, 30 Guilford Street, London WC1N 1EH, UK*

*Corresponding author at: Department of Clinical Neurosciences, University of Cambridge, R3 Neurosciences – Box 83, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK. Tel: +44-1223-217837; fax: +44-1223-217909.
E-mail address: eaw23@medschl.cam.ac.uk (E. A. Warburton).

Editorial Decision 5 February 2018; Accepted 14 February 2018

Abstract

Objective: Inner speech, or the ability to talk to yourself in your head, is one of the most ubiquitous phenomena of everyday experience. Recent years have seen growing interest in the role and function of inner speech in various typical and cognitively impaired populations. Although people vary in their ability to produce inner speech, there is currently no test battery which can be used to evaluate people's inner speech ability. Here we developed a test battery which can be used to evaluate individual differences in the ability to access the auditory word form internally.
Methods: We developed and standardized five tests: rhyme judgment of pictures and written words, homophone judgment of written words and non-words, and judgment of lexical stress of written words. The tasks were administered to adult healthy native British English speakers (age range 20–72, $n = 28$–97, varies between tests).
Results: In all tests, some items were excluded based on low success rates among participants, or documented regional variability in accent. Level of education, but not age, correlated with task performance for some of the tasks, and there were no gender difference in performance.
Conclusion: A process of standardization resulted in a battery of tests which can be used to assess natural variability of inner speech abilities among English speaking adults.

Keywords: Assessment; Language and language disorders; Test construction

Introduction

Inner speech, the ability to speak silently in your head, has been suggested to play an important role in memory (Baddeley & Hitch, 1974), reading (Corcoran, 1966), thinking (Sokolov & Onischenko, 1972), language acquisition (Vygotsky, Hanfmann, & Vakar, 1962), language comprehension (Blonskii, 1964), auditory hallucination (Brown, 2009), and even in consciousness and self-reflective activities (Morin & Michaud, 2007). Recent years have seen growing interest in the role and function of inner speech in various domains of investigation (Alderson-Day & Fernyhough, 2015; Morin, Runyan, & Brinthaupt, 2015), including cognitive language studies (Gauvin, Hartsuiker, & Huettig, 2013; Huettig & Hartsuiker, 2010), neuroscience (Gauvin, De Baene, Brass, & Hartsuiker, 2016; Geva, Jones, et al., 2011; Hurlburt, Alderson-Day, Kühn, & Fernyhough, 2016), child development (Lidstone, Meins, & Fernyhough, 2010) and aphasia (Geva, Bennett, Warburton, & Patterson, 2011; Langland-Hassan, Faries, Richardson, & Dietz, 2015), among others.

This interest resulted in the development of various tools aimed at evaluating inner speech abilities. With inner speech being essentially a phenomena which cannot be directly observed, many tools rely on individuals' reports of their own inner speech, using structured (e.g., Duncan & Cheyne, 1999; McCarthy-Jones & Fernyhough, 2011), or open-ended questionnaires (Morin, Uttl, & Hamper, 2011), or Descriptive Experience Sampling (Hurlburt, 1993; Hurlburt, Heavey, & Kelsey, 2013), where participants are asked to describe their inner experience at specific moments throughout the day. The advantages and disadvantages of those tools have been discussed elsewhere (Alderson-Day & Fernyhough, 2015; Uttl, Morin, & Hamper,

doi:10.1093/arclin/acy018

2011). Relevant to the current study, is the fact that all those tools describe inner speech as it is perceived by the individual, rather than objectively test the level of proficiency among participants.

Others have examined specific psycholinguistic aspects which have typically been examined in overt speech, in inner speech, using various experimental paradigms. For example, studies have looked at lexical bias (Nooteboom, 2005; Oppenheim & Dell, 2008), phonemic similarity effect (Oppenheim & Dell, 2008; Slevc & Ferreira, 2006), semantic similarity effect (Slevc & Ferreira, 2006), and verbal transformation effect (Sato et al., 2004) in inner speech. While these paradigms are helpful for testing specific hypotheses, they do not document people's ability to produce inner speech in general.

This study was therefore aimed at developing tests which measure inner speech as it is broadly described in the psycholinguistic literature. One of the cognitive models which gives vast attention to inner speech function is the working memory model by Baddeley and colleagues (Baddeley & Hitch, 1974; Buchsbaum & D'Esposito, 2008) which includes the phonological loop, a component which holds verbal information for a short period of time in which the information will decay unless actively rehearsed using the articulatory control process (Baddeley, 1966; Baddeley, Thomson, & Buchanan, 1975; Conrad, 1964). In Levelt's (1999) model of speech processing, the systems for language production and comprehension are described as partially separate – a view that has been supported by others (for example see Martin, 2003; Martin, Lesch, & Bartha, 1999). Levelt (1999) suggest that, for monitoring purposes, the phonological representation for production (phonological word) is perceived by the comprehension system, to create inner speech. Ozdemir, Roelofs, and Levelt (2007) reported that "uniqueness point", the place in the sequence of the word's phonemes at which it deviates from every other word in the language, influenced inner speech. As it is known that uniqueness point influences speech perception but not speech production, the authors concluded that, as initially suggested by Levelt, inner speech is dependent on the speech comprehension system. In contrast, Vigliocco and Hartsuiker (2002) argued that in the presence of overt speech, inner speech is processed by the production system alone. A study by Huettig and Hartsuiker (2010) showed differences in eye movements driven by internal speech compared to eye movements driven by one's own overt speech and overt speech produced by others. Although their findings support the idea that inner speech is not processed by the comprehension system, Vigliocco and Hartsuiker (2002) also suggest that in the absence of overt speech, what they describe as "conscious inner speech" can access the comprehension system. Therefore, according to Vigliocco and Hartsuiker (2002), inner speech can be processed either jointly by the production and comprehension systems, or by the production system alone, and the actual mechanism employed depends on the presence or absence of overt speech. Lastly, in connectionist models (for example, Dell & Oseaghdha, 1992; Seidenberg & McClelland, 1989), feedback connections within the production system can support the monitoring system without assuming two separate systems. All connections in the model are bidirectional, allowing the production of inner speech within this one system. This representation of one integrated system for both speech comprehension and production with bidirectional connections bare similarities to that proposed by Vigliocco and Hartsuiker (2002). This idea is supported by data from Marshall, Robson, Pring, and Chiat (1998) and others (for example, Nickels & Howard, 1995) who showed that in some cases of aphasia, comprehension can be relatively intact while monitoring of one's own errors is severely impaired and vice versa.

How much information is available in inner speech? Work by Oppenheim and Dell (2008) suggests that inner speech is phonetically impoverished in comparison to overt speech, because inner speech lacks some of the phonetic components present in overt speech, or, because the internal monitoring system fails to detect the full range of phonetic features of the produced inner speech. In addition, readers tend to vocalize more, as the difficulty of reading material increases, and have lower comprehension scores when prevented from vocalizing or even sub-vocalizing (Hardyck & Petrinov, 1970). However, Corcoran (1966) have shown that readers automatically access phonetics in inner speech.

In summary, inner speech has various functions, processing stages and access mechanisms (Perrone-Bertolotti, Rapin, Lachaux, Baciu, & Lœvenbruck, 2014; Postma, 2000). Different studies focus on different processing stages: from the initial stage in which thought is translated into words, to the final stage of articulation which may or may not accompany inner speech (Dell & Oppenheim, 2010). Accessing inner speech can be either spontaneous, as is the case when we think to ourselves in words, or stimuli driven, giving rise to inner speech which is produced, for example, during silent reading. Inner speech also supports error monitoring, and it is still debated whether this type of inner speech is based on the production system alone or on both the production and the comprehension components of the language processors.

In this study inner speech was defined as the ability to create an internal representation of the auditory word form, without producing overt sounds or articulation, and to apply computations or manipulations to this representation. Participants were asked to create inner speech based on a given stimuli. It is acknowledged that this definition of inner speech might relate to a different level of representation from the one used for error monitoring or spontaneous thinking. However, while it is a narrow definition, it is wide enough to include the various descriptions of inner speech as they appear in the psycholinguistic literature today. For this purpose, tests of inner speech abilities were developed. Here we describe these tests and their standardization process. We tested participants' ability to detect rhymes, homophones and the location of lexical stress in words and

non-words, using inner speech alone, elicited by a stimulus (written words or non-words, or pictures of familiar objects). All tasks have been used before in linguistic, psychological and neuro-imaging studies.

The use of the lexical stress assignment task was based on findings showing that adult speakers are able to determine and predict the location of lexical stress based on various linguistic parameters (Arciuli, Monaghan, & Seva, 2010; Jouravlev & Lupker, 2015). This knowledge about lexical stress in one owns language was previously exploited in various different tasks (Colombo, 1992; Kelly & Bock, 1988; Kelly, Morris, & Verrekia, 1998; Redford & Oh, 2016; Wade-Woolley & Heggie, 2015). In this study we asked participants to explicitly determine the location of the lexical stress in written words, varying in both length and stress location.

Homophones are words which have the same phonological form, that is, they sound exactly the same, but they are spelt differently. The existence of homophones in the English language has been exploited in various tasks (Davis & Herr, 2014; Gernsbacher & Faust, 1991; Kulczynski, Ilicic, & Baxter, 2017; Lukatela & Turvey, 1994; van Orden, 1987). Here we used a homophone judgment task, where participants are asked to determine whether two written words or non-words, which are spelt differently, sound the same. It has been shown previously that people access the phonological form of the word automatically (Lukatela & Turvey, 1994; van Orden, 1987), supporting fast and accurate response to homophone judgment tasks. Moreover, under conditions of cognitive load, people are less able to supress the multiple meanings of a homophone (for example, both "bye" and "buy" are activated when reading one of those words; Davis & Herr, 2014).

Lastly, the rhyme judgment task has been used extensively in the past. Early studies have shown that apart from retrieving a phonological word form, making a rhyme judgment requires working memory. This is not the case for homophony judgment (see early results by Brown, 1987; and reviews by Besner, 1987; Howard & Franklin, 1990). Much less work has been done on lexical stress, and to the best of our knowledge only one study evaluated whether lexical stress assignment requires working memory, concluding that, like rhyme judgment, it does (Campbell, Rosen, Solis-macias, & White, 1991). However, further work is required in this area. The ability to create and detect rhymes appears at an early age (Wood & Terrell, 1998) and is stable throughout adulthood (Geva et al., 2012; Zhuang, Johnson, Madden, Burke, & Diaz, 2016).

Methods

Participants

The word rhyme judgment task, word homophone judgment task and non-word homophone judgment task, adapted from the Psycholinguistic Assessments of Language Processing in Aphasia (PALPA; Kay, Coltheart, & Lesser, 1992), were completed by 63 participants (28M/35F; age range: 20–72, mean age: 41.7 ± 20.1; mean number of years of education: 15.6 ± 2.7).

The non-word homophone judgment task created by us (adapted non-word homophone task) was completed by 97 participants which included the 63 above participants (41M/53F/3 data not recorded; age range: 20–72, mean age: 38.1 ± 17.6; information regarding years of education was not recorded for the extra 34 participants).

A power calculation (Kirkwood & Sterne, 2003) relevant for the above tasks was based on preliminary results using similar tasks. The calculation showed that for required power of 80 (significance level of 0.05), a high expected success rate of 95% (as most tests used were originally developed for clinical use), and a relatively small difference to be detect (expected between 5% and 10%, based on existing data, and therefore set arbitrarily at 8%), a sample size of 59 participants will be required.

As part of the same study, the lexical stress task was completed by 28 participants (different from above, 16M/12F; age range: 21–63, mean age: 32.9 ± 10.4; information regarding years of education was not recorded).

The picture rhyme judgment task was completed as part of a follow-up study, conducted later. It was completed by a different group of 31 participants (12M/19F; age range: 22–71, mean age: 48.9 ± 20.0; mean number of years of education: 16.3 ± 2.9).

Participants were asked a series of questions regarding relevant medical history. Questions referred to neurological disorders (including stroke, Transient Ischemic Attack (TIA), and other neurological conditions), psychiatric conditions (with emphasis on those involving auditory hallucinations), developmental and learning difficulties (including dyslexia, specific language impairments, autistic spectrum disorder, ADHD, genetic disorders), and hearing and vision impairments. According to those self-reports participants had no history of neurological, psychiatric or language disorders, or any other documented learning difficulties. All had normal hearing, and normal or corrected vision. In addition, participants filled in a questionnaire regarding their language proficiencies (level of proficiency in writing, reading, speaking and understanding, in each of the languages they know). All reported to be native monolingual speakers of British English (note that many learnt a second

language in school, mostly French or German, but all described their level of proficiency as medium and below, in both production and comprehension). Information about ethnicity and socio-economic status was not collected. The study was approved by the Cambridge Research Ethics Committee and all participants read an information sheet and gave written informed consent.

Materials

Word rhyme judgment. The word pairs were taken from the PALPA (Kay et al., 1992). Participants were asked to determine whether two written words rhyme. For example, "bear" and "chair" rhyme, while "food" and "blood" do not. The test had altogether 60 pairs, divided into two lists. Half of the rhyming pairs and half of the non-rhyming pairs had orthographically similar ending (e.g., town – gown), while the other half had orthographically dissimilar ending (e.g., chair – bear). This way the test could not be solved successfully based on orthography alone, ensuring that the participants had to use their inner speech to solve the task.

Word homophone judgment. The word pairs were taken from the PALPA (Kay et al., 1992). Participants had to determine whether two words sound the same. That is, whether the words are homophones. For example, "might" and "mite" are homophones, while "ear" and "oar" are not. This test had 40 pairs of words, divided into two lists.

The pairs in both tasks were divided into the two lists for ease of administration. The frequency of the words was calculated, in order to evaluate whether the lists varied in word frequency. Lemma frequencies for the combined spoken and written form were taken from the Celex (Baayen, Piepenbrock, & Gulikers, 1995). When a word on the list corresponded to two different lemmas (for example, rush can be both a noun and a verb), the combined frequency was calculated. Overall, the four lists did not differ in word frequency (one-way ANOVA, $F_{(182,3)} = 0.61$, $p = .61$).

Non-word homophone judgment (PALPA and adapted). In this task participants had to determine whether two letter strings sound the same. That is, whether the non-words are homophones. For example, "zole" and "zoal" is a pair of non-word homophones, while "hane" and "hine" is a pair of non-words which are not homophones. 20 of those pairs were taken from the PALPA (Kay et al., 1992), half of which are homophones and half not. In the PALPA, the homophone and non-homophone pairs are matched one-to-one for visual similarity (Kay et al., 1992). Since this number of test items was small, 40 additional pairs of non-words were developed in a similar manner, i.e., homophone and non-homophone pairs were matched one-to-one for visual similarity. Each pair of non-words was created by replacing consonants or vowels in an existing English word, therefore creating two non-words. This was done to ensure that the non-words followed English spelling rules. As in the previous tests, this test could not be successfully solved based on orthography alone. See Supplementary material for the list of stimuli.

Lexical stress. Participants were asked to indicate where the stress was in each of 78 different words. We randomly selected words from the Oxford Dictionary, which varied in length, ranging from 2 to 5 syllables. In order to prevent participants from indicating stress location based on the word length alone, in each category of word length we included words which varied in their stress location. We selected frequent words (according to the Celex; Baayen et al., 1995), and aimed at having similar number of words in the different categories (different number of syllables, different stressed syllable). Stimuli parameters are presented in Table 1. See Supplementary material for the list of stimuli.

Table 1. Stimuli used for the lexical stress task

Number of syllables (Total number of items)	Stress location	Number of items	Example
2 (29)	first	16	Under
	second	13	Decide
3 (27)	first	11	Cinema
	second	9	Banana
	third	7	Volunteer
4 (20)	first	4	Supermarket
	second	9	Activity
	third	7	Conversation
5 (2)	third	2	University

Picture rhyme judgment. To create stimuli for the picture rhyme judgment task, 130 nouns were chosen, which created pairs of rhymes according to the Oxford British Rhyming Dictionary (Upton & Upton, 2004). Words were chosen if both words of a rhyming pair had corresponding black and white photos. All photos were taken from a large database of photos used extensively in the past for language studies, by our group and others. The words' endings in each pair differed in their orthography, so that participants could not make the rhyme judgment based on orthography alone. For each noun, a black and white photo was used (360 × 362 pixels, white background). Thirty native British-English speakers (14M/16F; age range: 23–63, mean age: 35.7 ± 12.7; mean number of years of education: 18.1 ± 3.8) were asked to name the pictures using one word. Pictures were presented in four blocks and the order of blocks was counterbalanced between participants. Pictures with high naming agreement (≥95%) were chosen for the task.

The final list contained 36 word pairs, out of which 26 of the pairs rhymed and 10 did not rhyme. A potential methodological problem was that if each picture appeared once in a rhyming pair and once in a non-rhyming pair, participants could develop a strategy whereby they might not use inner speech for some trials, but rather remember that if they saw the picture before in a rhyming pair, it is necessarily a non-rhyming pair this time, and vice versa. To avoid this, some pictures appeared only in rhyming pairs (once or twice) and some pictures appeared once in a rhyming pair and once in a non-rhyming pair. See Supplementary material for the list of stimuli.

Procedure

Word rhyme, word homophone and non-word homophone judgment tasks. Stimuli were presented on a paper and participants were asked to tick a YES column, if the words rhyme or are homophones, or tick a NO column, if the words do not rhyme, or are not homophones (see adapted non-word homophone judgment task in the Supplementary material, for example of the stimuli presentation). Participants first performed each task using inner speech alone and immediately afterwards were asked to read all the words aloud. This allowed recording the natural variability in the pronunciation of the words. The order of the tasks was randomized between participants.

Lexical stress. Lexical stress was defined to participants as 'some part of a word which is stronger, louder or longer, than others. The stressed part of the word gets the most emphasis'.

Participants were then given the following four examples: (1.) In the sentence: "You are the main suspect!", the stress in the word "suspect" is on the first syllable: suspect. (2.) In the sentence: "I suspect you!", on the other hand, the stress in the word "suspect" is on the last syllable: suspect. (3.) In the word "kingdom", the stress is on the first syllable: kingdom. (4.) In the word "Japan", the stress is on the last syllable: Japan. More examples were given if needed. In the task, words appeared twice. On the left, the word was printed as a whole, and participants were advised to read the word on this column first, and make their judgment. On the right hand side the word appeared segmented into syllables, and participants were asked to circle the part of the word that is stressed.

Picture rhyme judgment. Participants were first shown the pictures and asked to name them. When a naming error occurred the correct name was given, and participants were asked to name those pictures again until named using the desired word. This procedure was employed by others in their picture naming studies and it has the advantage of ensuring that participants produce the desired words during the performance of the task. As in the tasks above, variations in pronunciation, including regional accent, were recorded but not corrected. In each trial participants were presented with two pictures and had to indicate whether the words rhyme or not, by pressing one of two buttons, with their left hand. In each trial, the words "yes" and "no", together with a "v" and an "x", respectively, appeared at the bottom of the screen to remind participants that the left button corresponds to "yes" and the right button corresponds to "no".

In all tasks participants were instructed to use only inner speech, avoiding vocalization or articulation movements. Before each task participants practiced the task using examples, and the experimenter insured that the participant is not producing any overt speech. All tasks were administered by either one of the authors (SG) or a research assistant (SB).

Data Analysis

Scoring of the inner speech rhyme and homophone tasks was based on the judgment given to a word pair, with possible answers being correct or incorrect. Hence, every pair judged incorrectly was scored as one error. In the lexical stress task each word was scored as correct or incorrect. For each task the maximum possible raw score equals the number of included items (see Table 2). Here scores for all tasks are presented as ratio of correct responses (range 0–1).

Table 2. Items and participants included in each of the tasks

Task	Initial nu. of participants/items	Nu. of excluded participants/items	Final nu. of participants/items	Final nu. of items with "yes" response/total nu. of items (% items with "yes" response)
Word rhyme judgment	63/60	0/3	63/57	30/57 (52%)
Word homophone judgment	63/40	0/2	63/38	18/38 (47%)
Non-word homophone judgment	63/20	1/0	62/20	10/20 (50%)
Adapted non-word Homophone judgment	97/40	1/6	96/34	15/34 (44%)
Picture rhyme judgment	31/36	0/0	31/36	26/36 (72%)
Lexical stress	28/78	2/13	26/65	N/A

Item analyses were first performed, aimed at excluding any items on which group performance was not significantly above chance level. For the picture and word rhyme judgment, and the word and non-word homophone judgment tasks, possible responses were yes/no. Therefore, the threshold for exclusion was calculated based on the binomial distribution and items were excluded if the success rate was at chance level with $p > .05$. In the lexical stress task, items with success rates lower than 70% were excluded.

Next, data from participants who performed a specific task at chance level were excluded as well. For the yes/no tasks, this was based on the binomial distribution. For the lexical stress task, the chance level for the entire task was calculated based on the number of syllables in each word and the number of items with each number of syllables (for a word with two syllables, the probability of giving a correct answer by chance is 0.5; for a word with three syllables, the probability is 0.33; for a word with four syllables – 0.25; and for a word with five syllables – 0.20). The score not significantly different from chance was calculated to be 58%. This means that participants who scored 58% correct or lower were excluded.

Correlations between participants' age or amount of formal education (when available), and task performance, were examined using Kendall's Tau (two-tailed for correlation with age, one-tailed for correlation with education level), as Kolmogorov–Smirnov (K–S) tests indicated that the data deviated significantly from the normal distribution for all tests (K–S test, $p < .01$). We applied Benjamini–Hochberg FDR correction for multiple comparisons (Benjamini & Hochberg, 1995).

Results

Exclusion of Participants

One participant (70-year-old/M) performed at chance level on the non-word homophones judgment task (65% correct, $p = .074$); and another participant (44-year-old/F) performed at chance level on the adapted non-word homophone judgment task (62% correct, $p = .054$). Therefore, their data were excluded from these tasks. The data of two participants (22-year-old/M who scored 48% correct and 34-year-old/F who scored 46% correct), were excluded from the lexical stress analysis.

Exclusion of Items

Based on item analyses, the following items were excluded, as group performance did not significantly differ from chance: two items of the word rhyme judgment task (hush-bush, date-plait); one of the word homophone judgment task (dual-jewel); and three items of the adapted non-word homophone judgment task (shink-shynke, maccep-maxep, tousen-towsen). Based on low success rates, 13 items were excluded from the lexical stress task (one item with two syllables: hotel; eight items with three syllables: difficult, committee, understand, comprehend, elephant, photograph, contradict and hurricane; four items with four syllables: escalator, information, elevator and economic). Lastly, some further items were excluded based on documented regional variations in accent: one item from the word rhyme judgment task (gull-full); one item from the word homophone judgment task (bury-berry); and three items from the adapted non-word homophone judgment task (ama-amah, qaffal-kaphal, and thirm-theerm). A summary of the number of included items in each task is presented in Table 2.

Behavioral Performance

Table 3 presents the behavioral data for all the tasks used. The same 62 participants performed four of the tasks: word rhyme judgment (57 items), word homophone judgment (38 items), non-word homophone judgment (20 items) and adapted non-word homophone judgment (34 items). A multivariate test revealed that participants did not significantly differ in d' calculated for the

Table 3. Behavioral data for the different tasks

	Ratio correct					*d′*			
	Min	Max	Mean	Median	Std.	Min	Max	Median	Std.
Word rhyme judgment	0.77	1.00	0.96	0.96	0.039	−6.06	1.40	.296	1.478
Word homophone judgment	0.68	1.00	0.96	0.97	0.052	−9.36	1.22	.135	1.719
Non-word homophone judgment	0.70	1.00	0.95	1.00	0.073	−4.70	1.16	1.160	1.584
Adapted non-word homophone judgment	0.65	1.00	0.89	0.91	0.076	−3.64	2.00	0.244	1.268
Lexical stress	0.60	1.00	0.89	0.97	0.124			N/A	
Picture rhyme judgment	0.83	1.00	0.96	0.97	0.041	−4.49	1.33	0.552	1.533

different tasks ($F_{(3,60)} = 0.66$, $p = .58$), but did differ significantly in ratio correct ($F_{(3,60)} = 43.09$, $p < .001$). Planned comparisons using paired-sample *t*-test revealed that the adapted non-word homophone judgment task was significantly more difficult than the other three tasks ($t > 7$, $p < .001$ for all), and that the non-word homophone judgment task was significantly more difficult than the word homophone judgment task ($t_{(61)} = 3.03$, $p = .004$) and the word rhyme judgment task ($t_{(61)} = 2.01$, $p = .049$). See Fig. 1.

Similarly, the three tasks adapted from the PALPA significantly correlated with each other with regard to ratio of correct responses (word rhyme – word homophone Pearson's $r = 0.29$, $p = .023$; word rhyme – non-word homophone Pearson's $r = 0.29$, $p = .021$; word homophone – non-word homophone Pearson's $r = 0.69$, $p < .001$), while scores of the adapted non-word homophone judgment task significantly correlated only with the other two homophony tasks (significant correlation with word homophone task Pearson's $r = 0.33$, $p = .008$; and with non-word homophone task Pearson's $r = 0.57$, $p < .001$; non-significant correlation with word rhyme task Pearson's r = 0.10, $p = .432$).

Results were similar for *d′*: correlations were significant between the PALPA tasks (word rhyme – word homophone Pearson's $r = 0.30$, $p = .016$; word rhyme – non-word homophone Pearson's $r = 0.38$, $p = .003$; word homophone – non-word homophone Pearson's $r = 0.60$, $p < .001$), while d′ scores of the adapted non-word homophone judgment task significantly correlated only with the non-word homophone task (Pearson's $r = 0.47$, $p < .001$; non-significant correlation with homophone task Pearson's $r = 0.23$, $p = .074$; and with word rhyme task Pearson's $r = 0.9$, $p = .508$).

Age did not correlate with behavioral performance (ratio correct or *d′*) for any of the tasks (two-tailed Kendall's Tau, $p > 0.1$ for all). However, having more years of formal education correlated with better performance on the homophone tasks. This correlation was significant both for the word task (Kendall's Tau = 0.29, $p = .002$; Kendall's Tau = 0.28, $p = .003$; for ratio correct and *d′*, respectively) and for the non-word task (Kendall's Tau = 0.31, $p = .002$; Kendall's Tau = 0.28, $p = .003$; for ratio correct and *d′*, respectively), but not for the adapted task (Kendall's Tau, $p > .05$ for both measurements). Note that information regarding number of years of formal education was available for two of the three cohorts of participants (those performing the PALPA and adapted non-word homophone judgment tasks, and those performing the picture rhyme judgment task, reported below). However, there was no difference in education level between these two cohorts (independent sample *t*-test, $t_{(91)} = 1.07$; $p = .29$). There were no differences in performance between males and females on any of the tasks (*d′* or ratio correct, independent sample *t*-tests, $t < 2.0$, $p > .1$ for all tests).

For the word rhyme judgment task, we initially selected an equal number of items with similar and dissimilar ending, in order to evaluate whether participants rely on an orthography-based strategy for solving the task. A multivariate test (with the factors: orthographic ending, being "same" or "different"; and correct response, being "yes" or "no"), showed that there were no main effects of orthographic ending ($F_{(1,57)} = 0.14$, $p = .71$) or correct response ($F_{(1,57)} = 0.94$, $p = .34$). However, an interaction effect was significant ($F_{(1,57)} = 6.22$, $p = .016$), and post-hoc independent sample *t*-tests showed that when the two words in a pair had different orthographic endings, there was no significant difference in performance between rhyming and non-rhyming pairs ($t_{(26)} = 1.17$, $p = .25$). This means that participants were equally likely to judge correctly a pair like "bear-chair" (pair rhymes) and "pea-play" (pair does not rhyme). However, when the two words in the pair had the same orthographic ending participants made significantly more errors if the correct answer was "No" (i.e., the words do not rhyme) ($t_{(26)} = 2.16$, $p = .045$). This means that participants were more likely to make an error on pairs such as "food – blood" (similar orthographic ending, pair does not rhyme), compared to pairs such as "town-gown" (similar orthographic ending, pair does rhyme). See Fig. 2, where an example for relevant items is added to ease interpretation.

Lexical Stress

The distribution of items' success rate did not significantly deviate from the normal distribution (K–S test, $p > .05$). Therefore, parametric tests were used. The success rates for words with different number of syllables were compared. The group of words with five syllables was excluded from this analysis since it only had two items. The difference between success rates

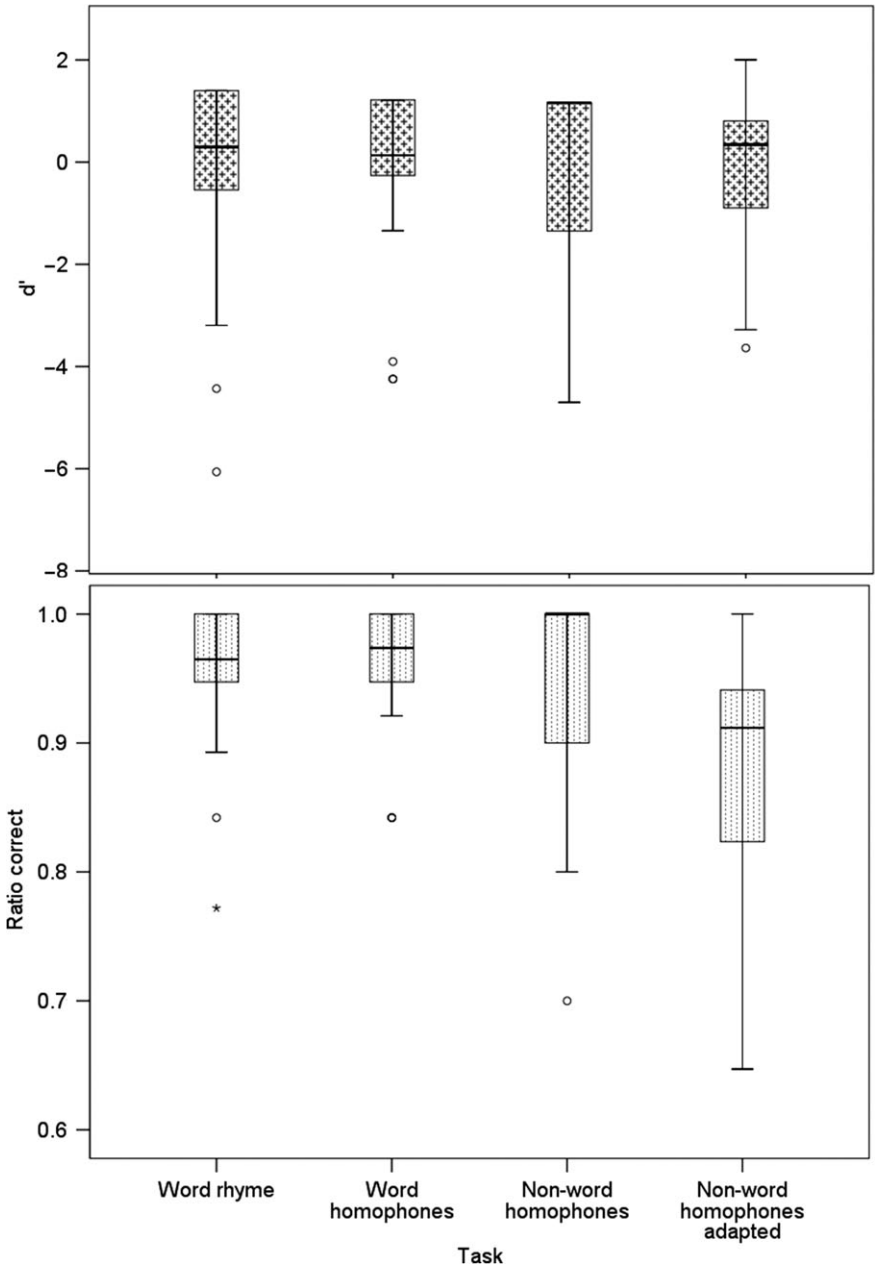

Fig. 1. Box plots representing d′ (top panel) and ratio correct (bottom panel) for the four tasks completed by 62 of the participants (word rhyme judgment, word homophone judgment, non-word homophone judgment and adapted non-word homophone judgment).

approached significance (one-way ANOVA, $F_{(2,60)} = 2.97$, $p = .059$). This suggests that the lexical stress of shorter words was more likely to be indicated correctly. Planned comparisons showed that while the words with two and three syllables differed significantly in their success rates (independent sample *t*-test, $t_{(48)} = 4.23$ $p < .001$), words with three and four syllables did not significantly differ from each other in their success rates (independent sample *t*-test, $t_{(45)} = 0.39$ $p = .695$). See Fig. 3. There were no differences in performance between males and females on the task (independent sample *t*-tests, $t_{(24)} = 0.53$, $p = .59$).

Picture Rhyme Judgment

Age or number of years of formal education did not correlate with behavioral performance (ratio correct or *d′*, Kendall's Tau, $p > .1$ for both). There were no differences in performance between males and females on either outcome measurements (*d′* or ratio correct, independent sample *t*-tests, $t < 1.0$, $p > .7$ for both).

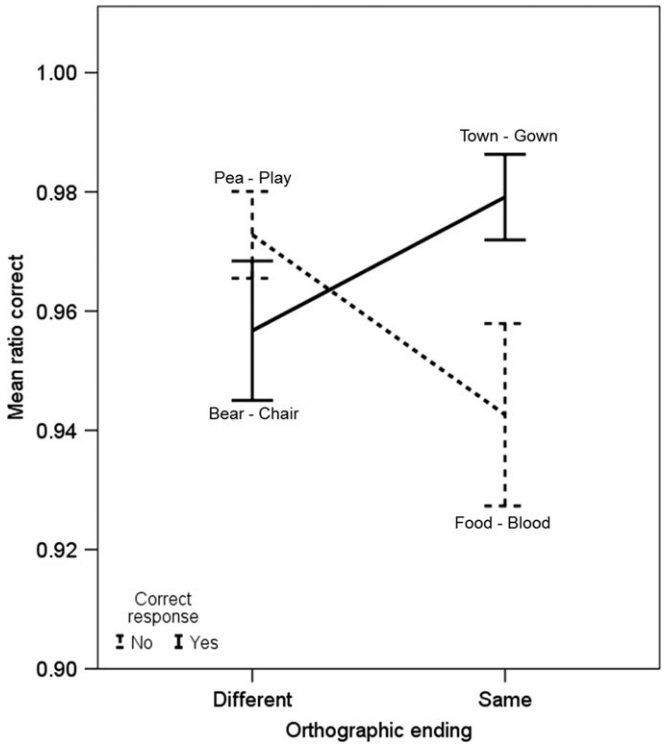

Fig. 2. Ratio correct according to item type, word rhyme judgment task (error bars represent standard error).

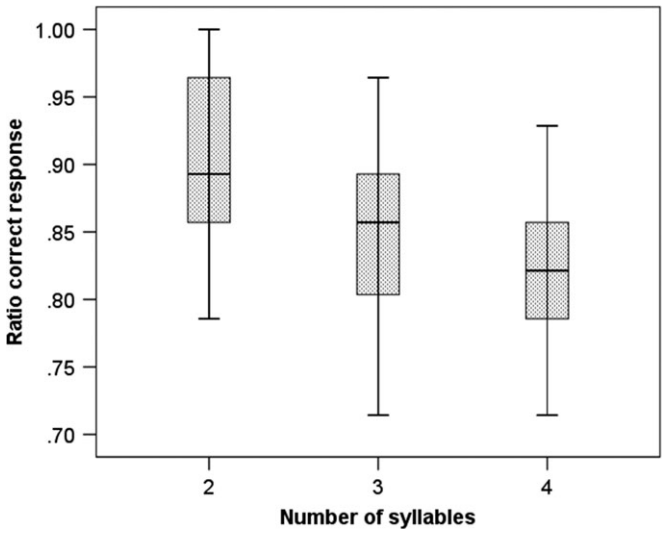

Fig. 3. Ratio correct for words with different numbers of syllables, lexical stress task.

Discussion

Standardization of the Tests

This study aimed to standardize tasks specifically designed to examine inner speech abilities. Native speakers of British English were tested and their performance was recorded. These scores can now be used to assess inner speech of native healthy English speakers.

Level of education, but not age, correlated with performance on some of the tests. The relation between language abilities and formal education is well documented. For example, Zanini, Bryan, De Luca, and Bava (2005) found that level of education was correlated with verbal communication ability, in a sample of Italian healthy adults. Similar results were found in a group of Mexican adults (Ardila, Ostrosky-Solis, Rosselli, & Gomez, 2000) and African and white Americans (Manly et al., 1998; Marcopulos, McLain, & Giuliano, 1997). The influence of education level on language ability can be complex and it interacts with other related variables such as socio-economic status (Dotson, Kitner-Triolo, Evans, & Zonderman, 2009), racial background (Dotson et al., 2009; Marcopulos et al., 1997), and literacy (Manly et al., 1999). Manly et al. (1999) suggest that it is the quality, rather than the amount of education that influences language ability. However, in the current study it was difficult to assess the quality of education participants received throughout their lives. In addition, the high level of education of some of the participants in the study might limit the external validity of those tests where performance correlates with level of education. This should be explored further in future studies.

Apart from a study by Meijer et al. (2008), the studies above did not find a correlation between age and language performance. Behavioral studies usually show that language production is affected by age, with word retrieval as the most affected domain (Ivnik, Malec, Smith, Tangalos, & Petersen, 1996; Marien, Mampaey, Vervaet, Saerens, & De Deyn, 1998; Mitrushina & Satz, 1995). However, this is more likely to be related to difficulties in lexical retrieval than to degradation of the phonological word form (Burke, Mackay, Worthley, & Wade, 1991; Heine, Ober, & Shenaut, 1999). The tasks used here minimized the demands on lexical retrieval (by either using written words or exposing participants to the desired object name in the picture task, before administering the task itself). Therefore, lack of age effects is consistent with previous studies.

Factors Contributing to Task Performance

Based on the definition of inner speech given in the Introduction, we argue that completion of all the tasks included require inner speech. And indeed, the correlation between performance scores (ratio correct and d') of the different tasks was highly significant in most cases. However, other cognitive processes and linguistic knowledge can contribute to task performance as well, and specifically explain the difference in performance levels between the tasks, especially when comparing the tasks which were completed by the same participants ($n = 62$). With regard to the rhyme judgment task, it is well established that silent rhyme judgment requires working memory (see Introduction), regardless of the nature of the stimuli (pictures or words). Therefore, performance on the task can be influenced by working memory abilities, not just inner speech abilities per se. This means that rhyme judgment tasks should not be used as measures of inner speech by themselves. This is true especially when studying clinical populations who suffer from verbal working memory impairments (e.g., Caspari, Parkinson, LaPointe, & Katz, 1998), or inner speech deficits (e.g., Feinberg, Gonzalez Rothi, & Heilman, 1986; Geva, Bennett, et al., 2011). Only by looking at convergent results from the various tasks used here researchers can reliably define participants' inner speech ability. Moreover, one can measure verbal working memory directly, using standardized tests such as digit span, to determine statistically to what extent task performance reflects inner speech abilities, and to what extent it reflects working memory capacity.

In addition, performance of the rhyming task can potentially be influenced by the level of similarity between the two words in the pair with regard to the orthographic ending. This was addressed here directly, by manipulating the stimuli and analyzing the difference between different types of word pairs. Specifically, stimuli were constructed to ensure that subjects retrieved the phonological form of the word and use their inner speech, rather than relying on orthography alone to perform the task. When reading words (aloud or silently), in some conditions readers have competition between an orthographic and a phonological route, because in English the phonology retrieved from semantics is not entirely consistent with the phonology retrieved from sublexical orthography (Coltheart, Curtis, Atkins, & Haller, 1993; Paap & Noel, 1991; Plaut, McClelland, Seidenberg, & Patterson, 1996). Between the four conditions (orthographic ending: similar vs. different, correct response: rhyme vs. non-rhyme), the competition is greater when orthography and phonology are not congruent. And indeed, when the two words in the pair had the same orthographic ending participants made significantly more errors if the correct answer was "No" (i.e., the words do not rhyme, and there is competition between orthography and phonology, e.g., food – blood), compared to the relatively easy condition, where there is no competition between orthography and phonology (e.g., town – gown). When the two words in a pair had different orthographic endings, participants were equally good in giving the correct response, whether the words rhyme (e.g., bear – chair) or not (e.g., pea – play). These results largely replicate previous studies (Damian & Bowers, 2010), and especially the seminal work by Seidenberg and Tanenhaus (1979) who found these effects even in an auditory rhyme judgment task.

Only one task, the picture rhyme judgment task, employed pictures rather than written words. Performing this task requires lexical retrieval, prior to the production of inner speech. In order to minimize the effect of lexical retrieval ability on task performance, participants were asked to name the pictures in advance. This also increases the likelihood that incorrect responses

are a result of difficulties with inner speech, rather than simply retrieving a different word than intended. Including a picture-based task allows us to generalize our results beyond the reading modality.

Lastly, the main difference between tasks in the level of success can be attributed to a lexicality effect: performance on tasks using words was significantly better than performance on tasks using non-words. While different models of language processing vary significantly in their accounts of the language system (see for example Coltheart et al., 1993; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Martin, Saffran, & Dell, 1996; Paap & Noel, 1991; Plaut et al., 1996), all models define levels of lexical/semantic support for phonology. As a result, performance on word tasks will always be better than performance on non-word tasks (unless performance on non-words is at ceiling). In general, these models can easily account for the data presented here since words receive activation from both semantics and phonology, while non-words receive only phonological activation (Coltheart et al., 1993; Martin, 2003; Paap & Noel, 1991; Seidenberg & McClelland, 1989). Moreover, the connections between the phonological units of a word are well practiced and are therefore stronger than those of non-words (Acheson & MacDonald, 2009). Having these extra sources of information increases the chance of retrieving the correct phonological word form and therefore reduces errors (Wilshire, 2008). Semantic influences on healthy inner speech are not previously documented, but, as hypothesized, were found in this study.

One would expect the rhyme judgment task to be harder than the homophone judgment task, as homophone judgment is largely automatic, and requires fewer processes than the rhyme judgment (see Introduction). However, this was not confirmed using our data. Both d' and error rate were similar for the two tasks. It might be that given the high success rate in both tasks, only response time (RT) would prove sensitive to the difference between the tasks. It should be noted that in order to compare such a sensitive variable as RT, one would need to match the stimuli based on various linguistic variables (such as word frequency and length). Here, words were matched for frequency. Matching stimuli on various parameters, under the constraints of the tasks, while still creating enough word pairs for the task to have enough power, is, however, difficult.

With regard to the lexical stress task, there is no consensus on whether the performance on this task requires working memory or not. In this study we found that performance on words with two syllables was significantly better than performance on words with three syllables (the difference between words with three and four syllables only showed a trend in this direction), a finding which could be explained as a working memory influence: it is more difficult to hold longer words in working memory and analyze them, compared to shorter ones. However, even a 4-syllable word is well within the span of normal working memory. It is important to note that as words get longer, the probability of giving a correct answer by chance decreases. Therefore, the length effect observed here can be a mere chance effect.

Another variable which can contribute to task performance is implicit knowledge of language rules or patterns. Stress assignment in English obeys specific rules. For example, the most frequent pattern is to stress the first syllable of a word with two syllables (Cutler & Norris, 1988). This general rule suffices to correctly determine the stress of 83% of all disyllabic English words in the CELEX database (Rastle & Coltheart, 2000). More importantly, it was previously found that speakers are sensitive to these rules. This was seen in two experiments; one examining stress assignment in non-word reading, and the other, analyzing reading latencies and errors in stress assignment when reading aloud real words (Rastle & Coltheart, 2000). However, the definition of regularity is by no mean under consensus. For example, while some scholars argue that lexical information influence stress assignment (for example whether the word is a verb or a noun, Kelly & Bock, 1988), others suggest that lexical information is not essential (Rastle & Coltheart, 2000).

In addition, there is a debate in the literature regarding the question of when speakers access the information about a specific word's stress. Rastle, Coltheart and colleagues (Coltheart et al., 1993; Rastle & Coltheart, 2000) developed one of the few reading models which aim at explaining how stress is determined by readers. They argue that stress is defined by rules, rather than being an information which is attached to each and every word. However, they suggest that rules can be derived from orthographic, as well as phonological information. If rules are derived from orthographic information, then theoretically, speakers can determine stress before accessing the phonological word form, i.e., without inner speech. If rules are derived from phonology, on the other hand, then an impoverished form of inner speech (without stress location) is a prerequisite for stress assignment. The authors do not opt for one of the options. Levelt (1999) on the other hand, argues that the word stress is specified during the level of morpho-phonological encoding. This process involves the retrieval of a few types of information; among them are the metrical shape of the word, specifying the stress pattern, and the segmental makeup, specifying the phonemes. Levelt (1999) further argues that stress is only specified when it is different from the language default. The various types of information are then combined to create the phonological word. This implies that stress pattern is available to the speaker before inner speech is produced. There is some evidence suggesting that people indeed have access to the metrical shape of the word in the absence of access to the word form itself. Barton (1971) examined 16 patients with aphasia, who were able to determine the number of syllables of unnamed words, above chance. Similarly, it has been found that patients with conduction aphasia performed better than patients with Wernicke's aphasia or anomia, in tasks requiring determining the syllabic length of words they could not retrieve (Goodglass, Kaplan, Weintraub, & Ackerman, 1976). In their seminal study

of tip-of-the-tongue states (TOT), Brown and McNeill (1966) asked students to guess the number of syllables of words they could not retrieve. Participants were correct in 60% of their guesses in the main study and 47% in a pilot investigation. These studies suggest that speakers have information regarding the metrical shape of the word even without retrieving the word form. However, there are no studies to date showing that speakers can access stress without accessing the word form. Examining TOT states, Rubin (1975) showed that when participants retrieve words which are related to the target word, these are likely to have stress patterns similar to the target word. This data can be interpreted as providing evidence that speakers have knowledge of the stress of words they cannot name. However, Rubin acknowledges that this result can also emerge from the fact that the related words are phonologically similar to the target words, and therefore have a similar stress pattern.

In summary, previous studies do not give clear answers as to how and when speakers access lexical stress, and whether stress can be accessed without accessing the word form. However, we argue that this task can still be used to asses inner speech for a number of reasons. Firstly, participants were explicitly instructed to use inner speech and all participants testified that this is how they solved the task. Secondly, participants took long time to complete the task, suggesting that their judgment was not based on implicit knowledge. Thirdly, and most importantly, the task included a similar number of words with various stress patterns. If participants use only their implicit knowledge of stress patterns to complete the task, they are likely to provide a wrong answer on many of the items.

One more factor which influences inner speech performance is articulation. It has long been debated whether inner speech without sub-articulation exists (reviewed in Geva, in press). In the past, some have argued that every production of inner speech is accompanied by motor activation of the articulatory musculature (Jacobson, 1930, 1932), even if this can only be traced using Electromyography (EMG), while others argued that inner speech can be produced without any motor involvement (Postma & Noordanus, 1996; Wyczoikowska, 1913). Since the introduction of brain imaging, numerous studies have examined motor activation during inner speech, producing mixed results (Basho, Palmer, Rubio, Wulfeck, & Muller, 2007; Geva, Jones, et al., 2011; Huang, Carr, & Cao, 2002). Recent studies suggest that articulation is more likely to occur when task demands are higher (Hardyck & Petrinov, 1970; Kell et al., 2017; McGuire et al., 1996). Here we did not attempt to completely supress sub-articulation during task performance, as it might be an integral and natural part of inner speech production. However, measuring sub-vocalization and sub-articulation using MEG, filming participants' lips (Frings et al., 2006; Simmonds, Leech, Collins, Redjep, & Wise, 2014) or recording potential overt responses during the covert condition (McGuire et al., 1996), can contribute to our understanding of the mechanisms underlying performance of these inner speech tasks.

Study Caveats

Variations in accent are an important aspect of performance for most of the tasks used here. This variable was not formally manipulated in this study, for example, by sampling from different regions of the UK. However, looking at the rhyme and homophone judgment tasks, one might notice that three out of the five pairs that were excluded (hush-bush, gull-full and bury-berry), have clear regional variability in the pronunciation of one or both words in the pair. For example, while an English speaker from the south of England might pronounce "hush" and "bush", or "gull" and "full" as non-rhymes, an English speaker from Yorkshire is more likely to pronounce these as rhyming pairs. The opposite is true for the words "burry" and "berry", which are usually pronounced as homophones by residents of the south of England but not by people from the north of England. When using these tasks one might consider asking participants to read all the words aloud after completion of the task using inner speech, as done here. This can allow the experimenter to exclude items which have accent variations in the study's specific population.

A second caveat is that the method used to calculate chance level for the lexical stress task is based on the assumption that speakers have no implicit knowledge of language patterns. However, as discussed above, this assumption is inaccurate. Hence, the probability of correctly determining stress location by chance might have to take into consideration the regularity of the stress patterns in each word in the list.

Other than in the lexical stress task, for many participants performance of these tasks was at ceiling. This result has a couple of implications. Firstly, scores which reflect accuracy of performance on most of these tasks are more suitable to verify that participants have normal inner speech, rather than for documenting subtle differences in inner speech ability. The exception is the lexical stress task where inter-subject variability was higher. Secondly, RTs were not collected for most tasks in this study. However, RT can be useful both for shedding light on which strategy people are using to perform the task, as well as to look at variability in performance among participants, in those instances where variability in success rate is low.

Lastly, the study has some major limitations in its characterization of the population that completed the tasks. Firstly, data about level of education is missing for some of the participants, and data about racial background and socio-economic status

was not collected. Secondly, participants were asked about any history of medical, psychological or psychiatric comorbidities, as well as any documented learning difficulties. However, some of the participants attended school many years ago and definitions and diagnostic criteria have changed substantially over the years. This, as well as the fact that comorbidities were determined based on self-report, rather than a formal examination, presents a major limitation with regard to the definition of the study population.

Future Directions

As discussed in the Introduction, some existing tools for evaluating inner speech use include structured or open-ended questionnaires, as well as Descriptive Experience Sampling (DES). Future studies should evaluate the relationship between participants' performance on the tasks used here, and their reports according to these existing tools. However, the relationship between the two types of measures is by no means obvious. For example, it might be the case that participants who tend to use inner speech spontaneously more often, are also better in monitoring their own inner speech in structured tasks as the ones used here. However, this is not necessarily so. It might also be the case that even individuals who do not tend to often use inner speech spontaneously, when requested to perform a specific task, can access it as successfully as those individuals who frequently engage in inner speech during the course of their daily activities.

In addition, it is interesting to understand the relationship between inner speech and executive functions which might influence the use of inner speech, including working memory and general monitoring behavior. This can be tested in future studies by evaluating participants' performance on executive functions using standardized tasks such as digit span or letter-number sequencing, for working memory, and go/no-go or Towers of London, for monitoring behavior.

Only little attention has been given to the role of inner speech and its impairments in acquired adult disorders affecting language function. In aphasia, it has been shown that inner speech can be affected by stroke to various degrees, and that inner and overt speech functions are not necessarily affected to the same level (Geva, Bennett et al. 2011; Langland-Hassan et al. 2015; Stark, Geva & Warburtun, 2017). Using the tasks developed here, future work can evaluate the relationship between inner and overt speech functions in healthy adults, potentially providing further evidence for the partial independence between the two functions in the human brain.

Lastly, advances in neuropsychological methods, such as the use of EMG, eye-tracking, and brain imaging, can shed light on the mechanisms underlying inner speech. Recent years have seen extensive interest in the neurobiology of inner speech, with imaging studies highlighting the involvement of language areas, including the left hemispheric Broca's area, supramarginal gyrus and angular gyrus, among others (Geva, in press). Notably, findings of imaging studies vary widely, and those differences are probably partly a result of the very wide range of tasks used. In addition, questions regarding the level of motor activation during inner speech, the existence of inhibition of overt response, support from working memory and executive function to inner speech production, among others, are still not fully answered. By combining the tasks developed here with such methods, those questions can potentially be answered, bringing us a step closer to understanding this complicated phenomenon.

Conclusions

We developed and standardized five tasks to assess inner speech abilities: rhyme judgment of words and pictures, homophone judgment of words and non-words, and lexical stress assignment. These tasks can be used as pen-and-pencil tasks, or adapted for presentation on a computer screen, to allow for collection of response time as well. While different tasks might require different cognitive abilities other than inner speech, together the tasks can give a profile of a participant's inner speech ability.

Supplementary Material

Supplementary material is available at *Archives of Clinical Neuropsychology* online.

Funding

This work was supported by the Pinsent-Darwin Fellowship; Wingate scholarship; The Cambridge Overseas Trust and B'nai Brith Scholarship to S.G. This work was further supported by the Biomedical Centre Grant (BMC) to Cambridge from the National Institute of Health Research Biomedical Research Centre (NIHR BMRC) to E.A.W.

Conflict of interest

None declared.

Acknowledgements

We thank Sophie Bennett for helping with testing.

References

Acheson, D. J., & MacDonald, M. C. (2009). Verbal working memory and language production: Common approaches to the serial ordering of verbal information. *Psychological Bulletin*, *135*, 50–68. https://doi.org/10.1037/a0014411.

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, *141*, 931–965. https://doi.org/10.1037/bul0000021.

Arciuli, J., Monaghan, P., & Seva, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. *Journal of Memory and Language*, *63*, 180–196. https://doi.org/10.1016/j.jml.2010.03.005.

Ardila, A., Ostrosky-Solis, F., Rosselli, M., & Gomez, C. (2000). Age-related cognitive decline during normal aging: The complex effect of education. *Archives of Clinical Neuropsychology*, *15*, 495–513.

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). CELEX2 – Linguistic Data Consortium. Retrieved October 4, 2017, from https://catalog.ldc.upenn.edu/ldc96l14.

Baddeley, A. (1966). Influence of acoustic and semantic similarity on long-term memory for word sequences. *Quarterly Journal of Experimental Psychology*, *18*, 302.

Baddeley, A., & Hitch, G. (1974). Working memory. In Bower G. H. (Ed.), *The psychology of learning and motivation: Advances in research and theory*, *Vol. 8*, pp. 47–89). New York: Academic Press.

Baddeley, A., Thomson, N., & Buchanan, M. (1975). Word length and structure of short-term-memory. *Journal of Verbal Learning and Verbal Behavior*, *14*, 575–589.

Barton, M. I. (1971). Recall of generic properties of words in aphasic patients. *Cortex*, *7*, 73–82.

Basho, S., Palmer, E. D., Rubio, M. A., Wulfeck, B., & Muller, R. A. (2007). Effects of generation mode in fMRI adaptations of semantic fluency: Paced production and overt speech. *Neuropsychologia*, *45*, 1697–1706.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. WileyRoyal Statistical Society. https://doi.org/10.2307/2346101.

Besner, D. (1987). Phonology, lexical access in reading, and articulatory suppression: A critical review. *The Quarterly Journal of Experimental Psychology Section A*, *39*, 467–478. https://doi.org/10.1080/14640748708401799.

Blonskii, P. P. (1964). *Memory and thought. In selected works in psychology*. Moscow: Prosveshchenie Press.

Brown, G. D. A. A. (1987). Phonological coding in rhyming and homophony judgement. *Acta Psychologica*, *65*, 247–262. https://doi.org/10.1016/0001-6918(87)90052-7.

Brown, J. W. (2009). Inner speech: Microgenetic concepts. *Aphasiology*, *23*, 531–543.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Behavior*, *5*, 325–337.

Buchsbaum, B. R., & D'Esposito, M. (2008). The search for the phonological store: From loop to convolution. *Journal of Cognitive Neuroscience*, *20*, 762–778.

Burke, D. M., Mackay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue—what causes word finding failures in young and older adults. *Journal of Memory and Language*, *30*, 542–579.

Campbell, R., Rosen, S., Solis-macias, V., & White, T. (1991). Stress in silent reading: Effects of concurrent articulation on the detection of syllabic stress patterns in written words in english speakers. *Language and Cognitive Processes*, *6*, 29–47. https://doi.org/10.1080/01690969108406937.

Caspari, I., Parkinson, S. R., LaPointe, L. L., & Katz, R. C. (1998). Working memory and aphasia. *Brain and Cognition*, *37*, 205–223.

Colombo, L. (1992). Lexical stress effect and its interaction with frequency in word pronunciation. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 987–1003. https://doi.org/10.1037/0096-1523.18.4.987.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud – Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589–608.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, *55*, 75–84.

Corcoran, D. W. J. (1966). An acoustic factor in letter cancellation. *Nature*, *210*, 658.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology-Human Perception and Performance*, *14*, 113–121.

Damian, M. F., & Bowers, J. S. (2010). Orthographic effects in rhyme monitoring tasks: Are they automatic? *European Journal of Cognitive Psychology*, *22*, 106–116. https://doi.org/10.1080/09541440902734263.

Davis, D. F., & Herr, P. M. (2014). From bye to buy: Homophones as a phonological route to priming. *Journal of Consumer Research*, *40*, 1063–1077. https://doi.org/10.1086/673960.

Dell, G. S., & Oppenheim, G. M. (2010). Motor movement matters: The flexible abstractness of inner speech. *Memory & Cognition*, *38*, 1147–1160. https://doi.org/10.3758/MC.38.8.1147.

Dell, G. S., & Oseaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, *42*, 287–314.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*, 801–838.

Dotson, V. M., Kitner-Triolo, M. H., Evans, M. K., & Zonderman, A. B. (2009). Effects of race and socioeconomic status on the relative influence of education and literacy on cognitive functioning. *Journal of the International Neuropsychological Society*, *15*, 580–589. https://doi.org/10.1017/s1355617709090821.

Duncan, R. M., & Cheyne, J. A. (1999). Incidence and functions of self-reported private speech in young adults: A self-verbalization questionnaire. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *31*, 133–136. https://doi.org/10.1037/h0087081.

Feinberg, T. E., Gonzalez Rothi, L. J., & Heilman, K. M. (1986). "Inner speech" in conduction aphasia. *Archives of Neurology*, *43*, 591–593.

Frings, M., Dimitrova, A., Schorn, C. F., Elles, H. G., Hein-Kropp, C., Gizewski, E. R., et al. (2006). Cerebellar involvement in verb generation: An fMRI study. *Neuroscience Letters*, *409*, 19–23.

Gauvin, H. S., De Baene, W., Brass, M., & Hartsuiker, R. J. (2016). Conflict monitoring in speech processing: An fMRI study of error detection in speech production and perception. *NeuroImage*, *126*, 96–105. https://doi.org/10.1016/j.neuroimage.2015.11.037.

Gauvin, H. S., Hartsuiker, R. J., & Huettig, F. (2013). Speech monitoring and phonologically-mediated eye gaze in language perception and production: A comparison using printed word eye-tracking. *Frontiers in Human Neuroscience*, *7*, 818. https://doi.org/10.3389/fnhum.2013.00818.

Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *17*, 245–262. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1827830.

Geva, S. (n.d.). Inner speech and mental imagery: A neuroscientific perspective. In Langland-Hassan P., & Vicente A. (Eds.), *Inner Speech: New Voices*. Oxford, UK: Oxford University Press.

Geva, S., Bennett, S., Warburton, E. A., & Patterson, K. (2011). Discrepancy between inner and overt speech: Implications for post stroke aphasia and normal language processing. *Aphasiology*, *25*, 323–343. https://doi.org/10.1080/02687038.2010.511236.

Geva, S., Jones, P. S., Crinion, J. T., Price, C. J., Baron, J. C., & Warburton, E. A. (2011). The neural correlates of inner speech defined by voxel-based lesion-symptom mapping. *Brain*, *134*, 3071–3082. https://doi.org/10.1093/brain/awr232.

Geva, S., Jones, P. S., Crinion, J. T., Price, C. J., Baron, J.-C., & Warburton, E. A. (2012). The effect of aging on the neural correlates of phonological word retrieval. *Journal of Cognitive Neuroscience*, *24*, 2135–2146. https://doi.org/10.1162/jocn_a_00278.

Goodglass, H., Kaplan, E., Weintraub, S., & Ackerman, N. (1976). Tip-of-tongue phenomenon in aphasia. *Cortex*, *12*, 145–153.

Hardyck, C. D., & Petrinov, L. F. (1970). Subvocal speech and comprehension level as a ffunction of difficulty level of reading material. *Journal of Verbal Learning and Verbal Behavior*, *9*, 647–652.

Heine, M. K., Ober, B. A., & Shenaut, G. K. (1999). Naturally occurring and experimentally induced tip-of-the-tongue experiences in three adult age groups. *Psychology and Aging*, *14*, 445–457.

Howard, D., & Franklin, S. (1990). Memory without rehearsal. In Vallar G., & Shallice T. (Eds.), *Neuropsychological impairments of short-term memory* (pp. 287–318). Cambridge, UK: Cambridge University Press.

Huang, J., Carr, T. H., & Cao, Y. (2002). Comparing cortical activations for silent and overt speech using event-related fMRI. *Human Brain Mapping*, *15*, 39–53.

Huettig, F., & Hartsuiker, R. J. (2010). Listening to yourself is like listening to others: External, but not internal, verbal self-monitoring is based on speech perception. *Language and Cognitive Processes*, *25*, 347–374. https://doi.org/10.1080/01690960903046926.

Hurlburt, R. T. (1993). *Sampling inner experience in disturbed affect*. New York: Plenum. Retrieved from https://books.google.com/books?id=Z93VBQAAQBAJ&pgis=1.

Hurlburt, R. T., Alderson-Day, B., Kühn, S., & Fernyhough, C. (2016). Exploring the ecological validity of thinking on demand: Neural correlates of elicited vs. spontaneously occurring inner speech. *PloS One*, *11*, e0147932. https://doi.org/10.1371/journal.pone.0147932.

Hurlburt, R. T., Heavey, C. L., & Kelsey, J. M. (2013). Toward a phenomenology of inner speaking. *Consciousness and Cognition*, *22*, 1477–1494. https://doi.org/10.1016/j.concog.2013.10.003.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., & Petersen, R. C. (1996). Neuropsychological tests' norms above age 55: COWAT, BNT, MAE token, WRAT-R reading, AMNART, STROOP, TMT, and JLO. *The Clinical Neuropsychologist*, *10*, 262–278.

Jacobson, E. (1930). Electrical measurements of neuromuscular states during mental activities VII. Imagination, recollection and abstract thinking involving the speech musculature. *American Journal of Physiology-Legacy*. Retrieved from http://ajplegacy.physiology.org/content/91/2/567.abstract.

Jacobson, E. (1932). Electrophysiology of mental activities. *The American Journal of Psychology*. Retrieved from http://www.jstor.org/stable/1414531.

Jouravlev, O., & Lupker, S. J. (2015). Predicting stress patterns in an unpredictable stress language: The use of non-lexical sources of evidence for stress assignment in Russian. *Journal of Cognitive Psychology*, *27*, 944–966. https://doi.org/10.1080/20445911.2015.1058267.

Kay, J., Coltheart, M., & Lesser, R. (1992). *Psycholinguistic assessment of language processing*. Psychology Press.

Kell, C. A., Darquea, M., Behrens, M., Cordani, L., Keller, C., & Fuchs, S. (2017). Phonetic detail and lateralization of reading-related inner speech and of auditory and somatosensory feedback processing during overt reading. *Human Brain Mapping*, *38*, 493–508. https://doi.org/10.1002/hbm.23398.

Kelly, M. H., & Bock, J. K. (1988). Stress in time. *Journal of Experimental Psychology-Human Perception and Performance*, *14*, 389–403. https://doi.org/10.1037/0096-1523.14.3.389.

Kelly, M. H., Morris, J., & Verrekia, L. (1998). Orthographic cues to lexical stress: Effects on naming and lexical decision. *Memory & Cognition*, *26*, 822–832. https://doi.org/10.3758/BF03211401.

Kirkwood, B. R., & Sterne, J. A. C. (2003). *Essential medical statistics*. Oxford: Blackwell Science.

Kulczynski, A., Ilicic, J., & Baxter, S. M. (2017). Pictures are grate! Examining the effectiveness of pictorial-based homophones on consumer judgments. *International Journal of Research in Marketing*, *34*, 286–301. https://doi.org/10.1016/j.ijresmar.2016.07.002.

Langland-Hassan, P., Faries, F. R., Richardson, M. J., & Dietz, A. (2015). Inner speech deficits in people with aphasia. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00528.

Levelt, W. J. M. (1999). Models of word production. *Trends in Cognitive Sciences*, *3*, 223–232.

Lidstone, J. S. M. M., Meins, E., & Fernyhough, C. (2010). The roles of private speech and inner speech in planning during middle childhood: Evidence from a dual task paradigm. *Journal of Experimental Child Psychology*, *107*, 438–451. https://doi.org/10.1016/j.jecp.2010.06.002.

Lukatela, G., & Turvey, M. T. (1994). Visual lexical access is initially phonological: 1. Evidence from associative priming by words, homophones, and pseudohomophones. *Journal of Experimental Psychology. General*, *123*, 107–128. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8014609.

Manly, J. J., Jacobs, D. M., Sano, M., Bell, K., Merchant, C. A., Small, S. A., et al. (1998). Cognitive test performance among nondemented elderly African Americans and whites. *Neurology*, *50*, 1238–1245.
Manly, J. J., Jacobs, D. M., Sano, M., Bell, K., Merchant, C. A., Small, S. A., et al. (1999). Effect of literacy on neuropsychological test performance in non-demented, education-matched elders. *Journal of the International Neuropsychological Society*, *5*, 191–202.
Marcopulos, B. A., McLain, C. A., & Giuliano, A. J. (1997). Cognitive impairment or inadequate norms? A study of healthy, rural, older adults with limited education. *Clinical Neuropsychologist*, *11*, 111–131.
Marshall, J., Robson, J., Pring, T., & Chiat, S. (1998). Why does monitoring fail in jargon aphasia? Comprehension, judgment, and therapy evidence. *Brain and Language*, *63*, 79–107.
Marien, P., Mampaey, E., Vervaet, A., Saerens, J., & De Deyn, P. P. (1998). Normative data for the Boston Naming Test in native Dutch-speaking Belgian elderly. *Brain and Language*, *65*, 447–467.
Martin, R. C. (2003). Language processing: Functional organization and neuroanatomical basis. *Annual Review of Psychology*, *54*, 55–89.
Martin, R. C., Lesch, M. F., & Bartha, M. C. (1999). Independence of input and output phonology in word processing and short-term memory. *Journal of Memory and Language*, *41*, 3–29.
Martin, N., Saffran, E. M., & Dell, G. S. (1996). Recovery in deep dysphasia: Evidence for a relation between auditory-verbal STM capacity and lexical errors in repetition. *Brain and Language*, *52*, 83–113.
McCarthy-Jones, S., & Fernyhough, C. (2011). The varieties of inner speech: Links between quality of inner speech and psychopathological variables in a sample of young adults. *Consciousness and Cognition*, *20*, 1586–1593. https://doi.org/10.1016/j.concog.2011.08.005.
McGuire, P. K., Silbersweig, D. A., Murray, R. M., David, A. S., Frackowiak, R. S., & Frith, C. D. (1996). Functional anatomy of inner speech and auditory verbal imagery. *Psychological Medicine*, *26*, 29–38.
Meijer, W. A., de Groot, R. H. M., Van Boxtel, M. P. J., Van Gerven, P. W. M., & Jolles, J. (2008). Are age differences in verbal learning related to inter-stimulus interval and education? *Experimental Aging Research*, *34*, 323–339. https://doi.org/10.1080/03610730802273910.
Mitrushina, M., & Satz, P. (1995). Repeated testing of normal elderly with the Boston Naming Test. *Aging Clinical and Experimental Research*, *7*, 123–127.
Morin, A., & Michaud, J. (2007). Self-awareness and the left inferior frontal gyrus: Inner speech use during self-related processing. *Brain Research Bulletin*, *74*, 387–396. https://doi.org/10.1016/j.brainresbull.2007.06.013.
Morin, A., Runyan, J. D., & Brinthaupt, T. M. (2015). Editorial: Inner experiences: Theory, measurement, frequency, content, and functions. *Frontiers in Psychology*, *6*, 1758. https://doi.org/10.3389/fpsyg.2015.01758.
Morin, A., Uttl, B., & Hamper, B. (2011). Self-reported frequency, content, and functions of inner speech. *Procedia – Social and Behavioral Sciences*, *30*, 1714–1718. https://doi.org/10.1016/j.sbspro.2011.10.331.
Nickels, L., & Howard, D. (1995). Phonological errors in aphasic naming – Comprehension, monitoring and lexicality. *Cortex*, *31*, 209–237.
Nooteboom, S. G. (2005). Lexical bias revisited: Detecting, rejecting and repairing speech errors in inner speech. *Speech Communication*, *47*, 43–58. https://doi.org/10.1016/j.specom.2005.02.003.
Oppenheim, G. M., & Dell, G. S. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, *106*, 528–537. https://doi.org/10.1016/j.cognition.2007.02.006.
Ozdemir, R., Roelofs, A., & Levelt, W. J. M. (2007). Perceptual uniqueness point effects in monitoring internal speech. *Cognition*, *105*, 457–465. https://doi.org/10.1016/j.cognition.2006.10.006.
Paap, K. R., & Noel, R. W. (1991). Dual-route models of print to sound – Still a good horse race. *Psychological Research-Psychologische Forschung*, *53*, 13–24.
Perrone-Bertolotti, M., Rapin, L., Lachaux, J. P., Baciu, M., & Lœvenbruck, H. (2014). What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural Brain Research*, *261*, 220–239. https://doi.org/10.1016/j.bbr.2013.12.034.
Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.
Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, *77*, 97–131.
Postma, A., & Noordanus, C. (1996). Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Language and Speech*, *39*, 375–392.
Rastle, K., & Coltheart, M. (2000). Lexical and nonlexical print-to-sound translation of disyllabic words and nonwords. *Journal of Memory and Language*, *42*, 342–364.
Redford, M., & Oh, G. (2016). Children's abstraction and generalization of English lexical stress patterns. *Journal of Child Language*. Retrieved from https://www.cambridge.org/core/journals/journal-of-child-language/article/childrens-abstraction-and-generalization-of-english-lexical-stress-patterns/D70BA53946AD6C0919AE60A233768038.
Rubin, D. C. (1975). Within word structure in the tip-of-the-tongue phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *14*, 392–397.
Sato, M., Baciu, M., Loevenbruck, H., Schwartz, J. L., Cathiard, M. A., Segebarth, C., et al. (2004). Multistable representation of speech forms: A functional MRI study of verbal transformations. *NeuroImage*, *23*, 1143–1151. https://doi.org/10.1016/j.neuroimage.2004.07.055.
Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.
Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 546–554. https://doi.org/10.1037/0278-7393.5.6.546.
Simmonds, A. J., Leech, R., Collins, C., Redjep, O., & Wise, R. J. S. (2014). Sensory-motor integration during speech production localizes to both left and right plana temporale. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*, 12963–12972. https://doi.org/10.1523/JNEUROSCI.0336-14.2014.
Slevc, L. R., & Ferreira, V. S. (2006). Halting in single word production: A test of the perceptual loop theory of speech monitoring. *Journal of Memory and Language*, *54*, 515–540. https://doi.org/10.1016/j.jml.2005.11.002.
Sokolov, A. N., & Onischenko, G. T. (1972). *Inner speech and thought*. D. B. Lindsley, (Ed.). New York, London: Plenum Press.
Stark, B. C., Geva, S., & Warburtun, E. A. (2017). Inner speech's relationship with overt speech in Poststroke aphasia. *Journal of Speech, Language and Hearing Research*, *60*, 2406–2415.

Upton, C., & Upton, E. (2004). *Oxford rhyming dictionary*. Oxford, UK: Oxford University Press.

Uttl, B., Morin, A., & Hamper, B. (2011). Are inner speech self-report questionnaires reliable and valid? *Procedia – Social and Behavioral Sciences*, *30*, 1719–1723. https://doi.org/10.1016/j.sbspro.2011.10.332.

van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory & Cognition*, *15*, 181–198. https://doi.org/10.3758/BF03197716.

Vigliocco, G., & Hartsuiker, R. J. (2002). The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin*, *128*, 442–472. https://doi.org/10.1037//0033-2909.128.3.442.

Vygotsky, L. S., Hanfmann, E., & Vakar, G. (1962). Thought and language. In Hanfmann E., & Vakar G. (Eds.), *Studies in communication*. Cambridge, MA: M.I.T. Press.

Wade-Woolley, L., & Heggie, L. (2015). Implicit knowledge of word stress and derivational morphology guides skilled readers' decoding of multisyllabic words. *Scientific Studies of Reading*, *19*, 21–30. https://doi.org/10.1080/10888438.2014.947647.

Wilshire, C. E. (2008). Cognitive neuropsychological approaches to word production in aphasia: Beyond boxes and arrows. *Aphasiology*, *22*, 1019–1053. https://doi.org/10.1080/02687030701536016.

Wood, C., & Terrell, C. (1998). Preschool phonological awareness and subsequent literacy development. *Educational Psychology*, *18*, 253–274. https://doi.org/10.1080/0144341980180301.

Wyczoikowska, A. (1913). Theoretical and experimental studies in the mechanism of speech. *Psychological Review*, *20*, 448–458. https://doi.org/10.1037/h0076098.

Zanini, S., Bryan, K., De Luca, G., & Bava, A. (2005). The effects of age and education on pragmatic features of verbal communication: Evidence from the Italian version of the Right Hemisphere Language Battery (I-RHLB). *Aphasiology*, *19*, 1107–1133. https://doi.org/10.1080/02687030500268977.

Zhuang, J., Johnson, M. A., Madden, D. J., Burke, D. M., & Diaz, M. T. (2016). Age-related differences in resolving semantic and phonological competition during receptive language tasks. *Neuropsychologia*, *93*, 189–199. https://doi.org/10.1016/j.neuropsychologia.2016.10.016.